数据论文的内容规范性研究

■ 黄国彬 郑霞
北京师范大学政府管理学院 北京 100875

摘要：[目的/意义]科学数据的规范化管理、引用和重用已受到学术界的广泛关注。在此背景下，用于促进科学数据合理使用的数据论文、数据期刊大量涌现。然而，数据论文目前尚未形成统一化、标准化的体例。本研究归纳和提炼当前在数据出版界较为活跃的数据期刊所发表的数据论文的内容框架和核心模块，以期为相关人员或机构在数据论文的撰写、使用和管理提供参考。[方法/过程]综合应用内容分析法和比较研究法，依据科学数据使用者查找、获取与重用科学数据的思维模式，基于不同数据期刊发布的投稿指南或写作须知，从主题相关性、数据质量、数据产生及获取方法、应用场景、使用方法、补充说明6个维度剖析数据论文的内容框架和核心模块。[结果/结论]目前并没有一种数据期刊的数据论文投稿指南或写作须知能够全部涵盖数据论文的所有模块；数据论文内容构成与传统学术论文的内容构成既有联系又有区别；数据论文内容框架的必备模块侧重于描述科学数据的突出特性；当前尚未形成体现科学数据特性的规范化描述框架。

关键词：数据论文 科学数据 内容框架 内容规范
分类号：G239.2
DOI：10.13266/j.issn.0252-3116.2019.22.014

1 引言

在以数据为主力驱动的大数据时代，科研人员不再只是关注科研成果的结论，同时也重视作为结论的核心支撑的科学数据[1]。从专业角度追溯科学数据的产生过程，核查科学数据的合理性和严谨性，成为科研人员判定某一科研成果主要结论的可靠性以至决定是否引用、借鉴该科研成果时的首要意识。同时，科研人员在使用科学数据时也不再拘泥于其是否为本团队所创建，而开始尝试重用（reuse 或 secondary use）现有的由他人创建或提供的科学数据，探究从不同维度利用这些科学数据的新途径、新场景[2]。但是，对于没有参与到科学数据获取过程的科学数据重用者而言，直接从他人所提供的，涉及实验设计、仪器设备、试剂标本、研究方法、问卷调查、观察探测、仿真模拟等复杂因素的，缺乏相关的语法、语义和语用信息支持的原始科学数据中去把握这些原始科学数据的核心内容和质量价值，往往困难重重。因此，创建可为科学数据的产生与创建、应用场景、操作要领和局限性等加以介绍和说明的科学数据使用文档显得至关重要。在此背景下，数据论文（data paper）逐渐产生和兴起。

目前学界对数据论文的概念并未形成定论，但从其内容和功能上看，数据论文通常包含数据集和对数据集进行描述的说明文档。其中，说明文档是对科学数据的获得方式、内容构成、应用场景和使用方法等进行介绍和描述以方便科学数据的阅读和使用的指导性文档。在一定程度上，数据论文可以为科研人员、科研机构、科研成果发表机构等在核查科研成果的研究过程、核心论据和主要结论等时，更为方便地阅读和理解作者所使用的科学数据，可促进科研成果集成出版平台更为高效地描述、揭示、存储、组织、检索和传播科学数据和以科学数据作为核心支撑的其他各类科研成果，也有利于科学数据使用主体更为准确全面地确定拟引用、重用的科学数据的质量与价值。

为进一步规范数据论文的格式和内容，目前已有数据论文出版平台发布了数据论文的投稿指南、写作须知和内容模板等。为此，笔者拟通过选取若干具有代表性的数据论文出版期刊或平台当前在数据论文方面的各类规范性文档，基于科学数据使用主体阅读和选用科学数据的思维逻辑，实证剖析数据论文的内容构成和核心要素，归纳其共性特点，以期更好地促进数

作者简介：黄国彬(ORCID:0000-0001-9059-8285)，副教授，硕士研究生导师，博士；郑霞(ORCID:0000-0002-9788-5823)，硕士研究生，通讯作者，E-mail:zhengxia199405@163.com。
收稿日期：2019-04-14 修回日期：2019-07-12 本文起止页码：129-140 本文责任编辑：徐健

据论文的内容规范化和使用高效化。

2 国内外研究现状

当前有关数据论文研究的相关论文并不多。国内多数学者将其称为“数据论文”，而国外学者在有关研究中以“data paper”命名。笔者以“数据论文”“数据期刊”为关键词在CNKI、Web of Science等数据库中检索相关文献。调研发现，目前，国内外对数据论文内容构成要素的研究主要集中在以下3个方面：

（1）简单介绍某一种数据期刊规定的数据论文的内容框架。学者刘凤红等[3-4]先后对《Data in brief》期刊和《Ecology》期刊的构成内容进行了分析，指出数据论文通常包含标题、摘要、作者、版权声明、研究方法、数据文件、数据记录、技术验证、用法说明等要素。

（2）在没有明确样本选择标准的情况下，以若干数据期刊为例，归纳数据论文的内容共性模块。田稷、陈晔夏[5]对3种数据期刊的内容要素从摘要、背景介绍、数据方法、数据集、数据说明、附加信息等加以剖析；黄如花、李楠[6]从7家数据出版期刊的出版政策出发，对数据论文的描述结构进行了剖析，认为其通常包括标题、作者、摘要、方法标识符、存储位置、参考文献等内容；李修[7]选择了7家数据期刊，从数据论文出版质量控制的角度对其结构标准进行了剖析，概括出6个共性要素：提供可访问的链接、提供简要说明、作者贡献说明、数据采集方法和工具、数据共享协议、数据动态存储要求；L. Candela[8]等调研了15家不同出版平台出版的数据期刊，指明数据论文对科学数据集的描述项目主要包括数据可用性、利益冲突、覆盖范围、数据格式、数据许可、作者贡献等；随后，学者M. Sandra[9]等在L. Candela研究结果的基础上对上述15家出版平台出版的数据期刊的投稿指南、写作指导和推荐模板进行了剖析，并提出一种通用的数据论文构成体例，包含标题、作者、作者机构、摘要、介绍、方法、数据集描述、图表说明、致谢、链接、参考文献在内的11个要素。

（3）从数据论文与传统学术论文的区别、数据论文与专门元数据的映射等角度研究数据期刊的组成要素。屈宝强、王凯[10]根据数据论文与传统学术性论文的区别，将数据论文的内容模块分为传统学术交流相关信息（如标题、作者、摘要、关键词等）和数据集相关信息（数据集基本信息和数据处理信息）。V. Chavan[11]等从标题、作者、摘要、介绍、覆盖范围、方法、项目描述和数据集描述等角度研究了数据论文与GBIF（全球生物多样性信息网络）元数据的映射关系。

分析国内外关于数据论文内容构成的已有研究发现：①从研究素材上看，多数文献仅是从数据期刊刊载的数据论文中提炼共性要素，而不是从数据期刊的数据论文投稿指南等指导性文档中去提取数据论文的内容框架，这往往会因某几篇数据论文未能完整代表某一数据期刊对数据论文的通用要求而致使结论的可靠性受影响；②从研究角度来看，当前研究主要按各个内容模块在数据论文文档中出现的先后顺序来论述，而没有从数据论文使用者在选用科学数据时评价标准的重要次序来剖析这些模块的内在关系，包括主题相关性、质量评估、应用场景和使用方法等；③从样本选择来看，尽管相关学者简要介绍了数据论文的核心构成，但样本选择的标准尚不明晰，数据期刊的代表性、学科覆盖范围的全面性等问题有待考证。基于此，笔者以当前发表的数据论文数量排名前10位的数据期刊为研究对象，依据数据使用者甄选、提取、复用、共享科学数据的逻辑流程，选取10种数据期刊的投稿指南、写作须知等规范文档，剖析当前主流的数据期刊在数据论文内容框架方面的共性特点。

3 研究方法与数据来源

笔者调研了目前汇集数据期刊总数最多（134种）的Dryad[12]平台，结合Web of Science、Elsevier Science Direct和Wiley Online Library等数据库的检索结果，广泛调研现有的数据期刊，剔除未提供数据论文投稿指南或其内容描述不完整的数据期刊。若不同数据期刊提供的数据论文投稿指南相似度较高，则仅选择其中刊载数据论文数量最多的数据期刊。最终选择刊载数据论文不少于30篇、数量排名前10的数据期刊所制定的投稿指南、写作须知等规范文档作为进一步分析的样本来源。这10种数据期刊依次是：《数据简述》（*Data in Brief*）[13]、《科学数据》（*Scientific Data*）[14]、《生态学》（*Ecology*）[15]、《BMC研究报告》（*BMC Research Note*）[16]、《数据》（*Data*）[17]、《中国科学数据》[18]、《F1000研究》（*F1000 Research*）[19]、《地球科学数据期刊》（*Geoscience Data Journal*）[20]、《开放心理学数据杂志》（*Journal of Open Psychology Data*）[21]和《生态学研究》（*Ecological Research*）[22]，期刊详情见表1。其中，完全数据期刊是指专门刊载数据论文的数据期刊，混合型数据期刊是指同时刊载数据论文和传统学术论文的数据期刊。在此基础上，采用内容分析法和比较研究法，从主题相关性、数据权威性、获得方法和

实验方法、应用场景、使用方法和补充说明 6 个角度，剖析当前数据论文的内容框架与核心要素。

表 1　收录数据论文数量前 10 位的数据期刊的基本信息

编号*	期刊名称	是否有明确的投稿指南	学科领域	所隶属的出版平台	数据期刊类型	收录的数据论文数量**
1	*Data in Brief*	是	全学科	Elsevier	完全数据期刊	3 404
2	*Scientific Data*	是	全学科	Nature Publishing Group	混合型数据期刊	702
3	*Ecology*	是	环境和生态学	Ecological Society of America	混合型数据期刊	179
4	*BMC Research Note*	是	所有科学及临床学科	Biomed Central	混合型数据期刊	161
5	*Data*	是	自然科学、医学、金融经济等	Multidisciplinary Digital Publishing Institute	混合型数据期刊	78
6	中国科学数据	是	全学科	中国科学院计算机网络信息中心	完全数据期刊	67
7	*F1000 Research*	是	生命科学和医学的基础科学、转化和临床研究	F1000	混合型数据期刊	48
8	*Geoscience Data Journal*	是	海洋学；大气和海洋化学；地表与地质学；水文学；地球化学；行星与空间科学等	Wiley	混合型数据期刊	42
9	*Journal of Open Psychology Data*	是	心理学	Ubiquity	完全数据期刊	31
10	*Ecological Research*	是	动物科学；行为科学；生态学；进化与发育生物学；林业；植物科学等	Wiley	混合型数据期刊	30

注：* 为节省篇复，本文后续部分仅用表 1 的数据期刊编号代替数据期刊名称

**截止到 2019 年 2 月底，各数据期刊所刊载的数据论文总数

4　数据论文投稿指南的框架构成与内容分析

4.1　总体的内容构成

针对本次研究所选择的 10 种数据期刊的投稿指南、写作须知等规范文档的内容特点，笔者设计了包括主题相关性、数据权威性、获得方法和实验方法、应用场景、使用方法和补充说明等在内的 6 个一级类目，用以全面涵盖出现在数据论文投稿指南、写作须知等规范文档中的全部 25 个子项。在此基础上，采用内容分析法，统计各子项在这 10 种数据期刊中的分布情况，形成表 2 的“10 种数据期刊投稿指南的内容模块统计”。从表 2 可以看出，目前这 10 种数据期刊的数据论文投稿指南对多个一级类目都有所涉及，但是，在应用场景方面，多数数据期刊对此仍未做出明确要求。

4.2　六大模块的剖析

数据论文是指导科学数据使用的说明性文档，是科学数据利用者在使用科学数据时的指南。从使用科学数据时的思维逻辑出发，除了基于特定论文所提供的线索直接定位到某种科学数据(集)，通常情况下，科学数据使用者对某一种科学数据(集)进行检索、浏览、引用或重用，首要考虑的原则是主题相关性，即科学数据是否与其所研究的学科、专业、方向、领域、论题密切相关，其次是科学数据的质量、获得方法和实验方法。在此基础上，对于准备重用科学数据的使用者而言，将会进一步考察科学数据的应用场景、使用方法和相关的补充说明。因此，一份可用高效的数据论文，虽然会因不同数据期刊的编排风格导致其细分模块的出现次序与位置不同，但是，在内容的完备性上，综合考虑并全面涵盖前述这六个基本模块，将是数据论文能够更好地满足科学数据使用者阅读数据论文需求的重要保障。

4.2.1　主题相关性　主题相关性用于揭示数据论文所描述的科学数据涉及的学科、专业、方向、领域或话题，是科学数据使用者使用科学数据的首要考虑因素。对科学数据所涉及的主题描述是否准确、充分，直接关系到科学数据存储机构和发布平台对科学数据的有序组织并提供高效检索，也对科学数据使用者和管理者在科学数据的查找、使用和评价等方面有着直接的影响。目前，所调研的 10 家数据期刊，其体现主题相关性的细分模块主要涉及数据论文的标题、摘要、学科领域、数据集和关键词等。具体见表 3。

(1)标题。数据期刊投稿指南中规定的标题是指在总结提炼数据论文主要内容的基础上，得到的揭示数据主旨信息的浓缩式简短语句。通过快速浏览数据论文的标题，科研人员可以方便地判断科学数据的学科范围和利用价值。经分析，本次研究所选择的 10 种数据期刊均要求数据论文必备标题模块。不过，期刊 8 未对标题模块进行详细介绍，其他 9 种数据期刊均在投稿指南中对数据论文标题模块提出了具体要求。

表 2 10 种数据期刊投稿指南的内容模块统计

核心模块	内容构成＼期刊编号	1	2	3	4	5	6	7	8	9	10
主题相关性	标题	√	√	√	√	√	√	√	√	√	√
	摘要	√	√	√	√	√	√	√	√	√	√
	学科范围	√									
	数据(集)	√		√	√	√	√		√	√	
	关键词	√		√	√	√	√	√	√	√	√
数据质量评估	数据论文作者	√	√	√	√	√	√	√	√	√	√
	数据(集)权属信息	√		√		√	√		√	√	
	数据(集)完整性						√				
	数据(集)优劣程度			√	√			√		√	
	资助信息	√		√	√	√	√	√			√
数据产生背景、获得方法和实验方法	数据产生背景和目的		√	√	√	√	√	√		√	
	实验设计、数据采集和处理方法	√	√	√	√	√	√	√	√	√	√
	数据样本描述	√		√			√			√	
应用场景	数据可用性		√	√	√		√		√		
	技术验证		√	√				√			
	数据记录		√								
	数据价值	√					√				
使用方法	数据(集)存储位置			√	√	√	√			√	
	数据结构描述	√	√	√	√	√	√	√	√		√
	数据格式、类型与时效性说明	√	√	√	√	√	√	√	√	√	√
	数据使用说明		√	√		√	√			√	
补充说明	补充信息	√	√		√		√	√		√	√
	作者贡献		√		√	√	√	√		√	√
	致谢	√	√	√	√	√	√	√	√	√	√
	参考文献	√	√		√	√	√	√	√	√	

表 3 10 种数据期刊对数据论文在主题相关性方面的写作要求

期刊编号＼核心模块	数据论文的标题	数据论文的摘要	学科范围	数据(集)	数据论文的关键词
1	标题须明确、专指,并建议包含“数据”或“数据集”等表达	提供明确数据信息,纯描述性,说明存储库名称和引用编号	一级学科,二级学科	简要描述数据信息,数据内容可视化,数据报告需源于实验或观测	未明确规定
2	不超过 110 个单词,尽量不使用缩略语,禁止使用冒号和括号	简短(不超过 170 个单词),无引用,简要描述实验、数据重用性	–	–	–
3	简明,不超过 120 个单词,不应使用分类号、数值系列名称	简短(不超过 350 个单词),介绍数据库基本信息,不可过度解释、推导	–	提供数据集标题或主题、数据研究目的、数据信息(如数据主题范围等)、数据使用价值	最多 12 个
4	清楚,简明	简短(不超过 200 个单词),减少缩写,无引用,必须包含写作目的和数据价值两部分	–	提供有关数据内容的纯描述性信息,数据集存储于公共存储库,不超过 500 个词	最多 10 个
5	简明、具体、切题,可区分数据来源,专业词汇需用全称	简短(不超过 200 个单词),独立成段,客观反映数据内容,不得提供研究结果、结论	–	数据集的标题、DOI、创建者、发布者和位置(如 URL)	最多 10 个
6	体现数据对象、空间和时间范围	格式上:不超过 500 个单词,无引用,不分段;内容上:简要介绍数据集实验方法、条件、基本信息	–	–	4 ~ 8 个,适度体现时间和空间跨度
7	简明、具体、清晰	简明(不超过 300 个单词),无引用,不得使用缩略语。	–	–	最多 8 个
8	未说明	简明(不超过 300 个单词)	–	数据集名称、数据中心、DOI	最多 5 个
9	反映数据本身	简短(约 100 个单词左右),描述数据内容、数据收集方法、存储数据途径以及数据重用价值	–	数据集或数据文件在存储库中的名称	未明确规定
10	简短、具体、准确,不应使用缩略语	简明(不超过 350 个单词),若有引用应置于摘要和关键词之间。	–	–	最多 10 个

从内容上看,多数数据期刊要求数据论文的标题应当简明扼要地反映数据本身,体现数据对象的来源、时间和空间范围等关键信息;从格式上来看,标题应当尽量减少或避免缩略语或多余标点符号的使用。除了这些共性要求之外,不同期刊对标题模块的内容、格式、注意事项等细节描述还存在一些差异。

(2)摘要。数据期刊投稿指南中规定的摘要指的是简要、确切地概述科学数据重要内容的短文。摘要是对标题进一步的详细说明。总体来看,当前刊载数据论文数量排名前10位的数据期刊均对摘要的内容和格式提出了细节说明。在内容方面,10种数据期刊投稿指南均规定,摘要部分仅是对数据内容、数据产生背景、实验条件和方法、数据价值等信息进行简要客观描述,不得包含以下内容:①对科学数据进行抽象推导、解释分析;②对新的科学研究或科学发现予以阐释;③介绍与数据集相关联的研究性论文。在格式方面,10种数据期刊投稿指南提到,摘要部分应独立成段。当然,不同数据期刊对摘要字数的要求存在差异(见表3)。此外,与其他数据期刊的规定不同,仅期刊10允许摘要部分引用参考文献,并要求将引用内容置于摘要和关键词之间。

(3)学科范围。数据期刊投稿指南中规定的学科范围指的是科学数据所归属的相对独立的知识门类,用以界定和区分科学数据所涉及的不同研究领域。分析发现,目前仅期刊1对数据论文所描述的科学数据涉及的学科范围进行规范化界定,要求在数据论文中提供科学数据的一级学科(如物理、化学、生物等学科)和二级学科(即更细化的学科范畴)的信息。同时,为了提供更高效的检索入口,期刊1还设置有学科"类别"(category),科学数据提供者在提交科学数据时可以在27个大类和238个二级类目中选择合适的类目予以标记,以促进科学数据的发现、共享和使用。

(4)数据(集)描述。数据期刊投稿指南中规定的数据集描述是数据期刊要求数据论文创作者在数据描述阶段提供的有关科学数据的内容、研究背景、存储路径及其他补充信息的细节描述。这是数据论文与传统学术性研究论文的区别之一。当前刊载数据论文数量排名前10位的数据期刊中,有7种数据期刊在其投稿指南中对数据集描述进行了说明,而期刊2、期刊7、期刊10尚未对此提出强制性要求(见表3)。分析可知,7种数据期刊关于数据集描述项目的有关规定具有相似之处,比如,均要求在数据集描述部分提供数据集标题、数据类型、数据内容、数据产生背景和处理方法的细节描述、与数据集关联的DOI(或其他唯一标识符)等内容。

(5)关键词。数据期刊投稿指南中规定的关键词指的是经过提取和凝炼的、可涵盖数据论文全文内容的关键性词汇,可用于精准、高效地检索和定位数据论文。笔者选取的10种数据期刊中除期刊2未在投稿指南中对关键词提出明确要求外,其他9种期刊都要求数据论文必须提供关键词。值得注意的是,尽管期刊1和期刊9要求数据论文需包含关键词,但未对关键词的数量做出强制性规定,而其余7种期刊均明确规定了关键词的数量(见表3)。同时,期刊6要求关键词可适度反映时间和空间跨度。

4.2.2 数据质量 在完成科学数据的主题相关性判断后,科研工作人员需要对这些科学数据进行质量评估,以进一步确定其所关注的某一科研论文在实验设计和研究结论上的可靠性,或者是进一步确定是否引用或重用该科学数据。经分析可知,多数数据期刊通过数据论文的作者、数据(集)的权属信息、数据(集)的完整性和数据(集)的优劣程度、资助信息5个模块来直接或间接地呈现科学数据的质量,见表4。

(1)数据论文的作者。数据论文作者是指创作数据论文的主要科研人员。笔者选取的10种数据期刊普遍要求提供数据论文的作者及其相关信息,包括作者姓名、电子邮件、所属单位、地址、邮编等,并指定一位作者作为数据论文的通讯作者加以标注。其中,期刊2对数据论文作者信息提出了更为细致的要求,除上述基本内容外,还提出最多可列举6位对数据论文做出同等贡献的作者和6位联合监督者,并通过尾注形式注明"这些作者共同为此项工作做出了贡献"或"这些作者共同监督了此项工作的实施"。而期刊5要求按照PubMed/MEDLINE标准格式的要求填写完整的数据论文作者地址信息(城市、邮政编码、州/省、国家和所有电子邮件地址)。

(2)数据(集)权属信息。数据期刊投稿指南规定的数据集权属信息是涉及科学数据集的收集、创建、出版、利用等各个关键环节的所有权归属信息。目前,笔者所选择的这10种数据期刊中,用于描述数据集权属信息的模块主要包括:数据(集)创建者、数据(集)出版商、数据(集)作者机构、数据(集)作者电子邮件、数据(集)合作者等。基本上,这10种数据期刊投稿指南均在不同程度上要求列出数据集的创建者、收集者、出版商等细节。而期刊1和期刊3专门提供了规范化的元数据描述项,类似的还有期刊6增设了"数据库(集)基本信息简介"等模块来明晰科学数据的权属信息。

表4 10种数据期刊对数据论文在揭示数据质量方面的写作要求

期刊编号＼核心模块	数据论文作者	数据(集)权属信息	数据(集)完整性	数据(集)的质量控制	资助信息
1	作者姓名、邮箱、所属单位	数据创建者、作者机构、通讯作者及电子邮件、数据创建合作者	-	-	资助机构名称、资助编号、基金会、赠款编号
2	作者姓名、工作单位、地址、邮编、国家,最多列举6位数据贡献者和6位联合监督者	-	-	-	-
3	作者姓名、地址。若住址变更需标注	与数据集相关的主要研究人员的姓名及地址	-	数据集使用历史、数据质量控制	资助人、资助机构、资助编号、名称及地址
4	所有作者的全名、机构地址和电子邮件(协作组包含在内)	-	-	数据收集过程中存在的问题(300个单词以内)	说明资助机构具体贡献
5	作者全名,作者所属单位填写格式遵循PubMed/MEDLINE标准格式	数据集创建者、出版者	-	-	资助人姓名、资助编号
6	作者的姓名、单位/机构、城市、邮编	数据作者、数据通信作者	数据库(集)组成(如组成部分、数据量)、数据时间范围	-	基金来源、名称、编码
7	作者机构地址、邮编和电子邮件,多中心合作形式的直接责任人	-	-	数据质量控制和验证过程、误差数据及精度、缺失、断点等情况	资助人的姓名、资助编号、获资助的个人
8	作者全名、部门/机构、城市、国家	数据集创建者、出版商	-	-	-
9	作者的全名、机构	数据集创建者、收集者、出版商	-	列出数据质量控制方法	-
10	作者姓名、机构、通讯地址	-	-	-	在致谢部分提供资助机构的全称

(3)数据(集)的完整性。数据(集)的完整性是对科学数据集在时间维、空间维、类型维和格式维等角度进行综合表述,以体现科学数据的全面性和连续性。可以说,数据(集)的完整性是确保其高质量的必要条件之一。然而,从当前刊载数据论文数量排名前10位的数据期刊所提供的投稿指南、写作须知等规范文档来看,多数数据期刊对数据(集)的完整性并未提出强制性的要求,仅有期刊6简要提及数据完整性的基本要求。其在"数据库(集)基本信息简介"模块中要求描述数据库(集)的组成情况,如分为哪些组成部分,每部分的数据量是多少、数据收集的时间跨度等。

(4)数据(集)的质量控制。数据(集)的质量控制主要涉及科学数据的使用记录、局限性与质量管理。经调研可以发现:①在数据集的使用记录方面,10种数据期刊中仅有期刊3指出应提供数据集的使用历史记录。具体来说,数据使用记录包括使用数据请求(记录请求使用该科学数据的人员姓名、使用目的以及科学数据的实际使用过程)、数据集更新记录(描述对数据集进行的所有更新操作)、其他用户的评论和疑问(其他用户发现的可疑或异常数据、在数据使用过程中遇到的限制、当前尚未解决的问题等);②在数据集的局限性说明方面,10种数据期刊中仅期刊4要求提供不超过300个单词的有关数据(集)局限性的说明,包括在数据收集过程中出现的问题(如样本量较小、数据陈旧等)。该刊认为,有关数据局限性的真实描述对数据的引用和复用至关重要。③数据质量控制与评估。期刊3、期刊7和期刊9都要求提供数据质量控制和评估的细节描述,如对异常值的识别和处理、参考标准的校对、误差数据及相关精度、连续性数据中的缺失断点等情况予以说明。

(5)资助信息。数据期刊投稿指南中所规定的资助信息指的是创建、产生数据(集)的科研项目在资助来源方面的细节性信息,包括资助机构、资助级别、资助金额等,可在一定程度上帮助科学数据使用者把握科学数据的权威性。通常情况下,更为高规格的资助机构,经其大规模投入资金所产生的科研数据,可靠性上相对更有保障。期刊1、期刊3、期刊6和期刊9均要求在数据论文中提供资助机构或基金会的名称、资助编号和赠款编号等。

4.2.3 数据产生背景、获得方法和实验方法 数据产

黄国彬，郑霞. 数据论文的内容规范性研究［J］. 图书情报工作，2019，63(22)：129－140.

生背景、获得方法和实验方法是指依托特定知识背景和研究内容获得和创建科学数据的过程记录。经调研，笔者所选择的 10 种数据期刊均在不同程度上对此加以规定，如表 5 所示：

表 5　10 种数据期刊对数据论文在描述数据产生背景和获得方法方面的写作要求

期刊编号 \ 核心模块	数据产生背景和目的	实验设计、数据采集和处理方法	数据样本描述
1	－	实验设计和数据收集方法，图表、代码	实验因素、实验特点、数据源地理位置
2	实验设计、生成数据背景，实验目的	数据产生步骤和程序、实验设计描述、数据采集和处理过程	－
3	数据产生背景、研究起止时间、研究目的	实验设计、抽样过程、数据统计、现场试验、计算机模拟	实验区概貌、栖息地、土壤地质地貌、气候特征
4	数据来源、研究目的	以独立文档描述数据收集和处理的方法、技术、质量控制	－
5	数据内容概述、实验方法、数据收集背景、数据价值	数据收集和处理的方法、许可声明	－
6	研究意义、研究背景资料、数据价值	试验设计、试验过程的方法、程序、算法和分析	样本来源、数据结构、地理位置
7	研究目标、背景、原理	数据详细说明、数据深加工原因及过程	－
8	－	试验设备和仪器	－
9	资料收集背景（100－250 个单词）	数据收集方法描述（100－200 个单词）、影响因素、项目数量、操作过程	人口基本信息、选择被试的方法和过程
10	－	试验设计、测量步骤、化学试剂和仪器	－

（1）数据的产生背景和目的。数据期刊投稿指南中规定的数据产生背景和目的模块，指的是对数据（集）产生的知识背景、研究目标等加以概述。经调研，有 6 种数据期刊要求提供数据产生的背景和目的信息。这些数据期刊在其投稿指南中均要求，应对数据产生背景、数据来源、数据潜在使用价值和实验目的等内容进行概括性描述。

（2）实验设计、数据采集和处理方法。数据期刊投稿指南中规定的实验设计、数据采集和处理方法模块，要求描述和介绍采集、生成和处理科学数据的各类方法和工具，是数据内容描述的基础和核心。鉴于该模块在数据论文内容构成中占据重要地位，10 种数据期刊都强制性要求数据论文提供数据处理的描述性信息。内容主要包括：实验设计、抽样方法、现场试验、计算机处理（如数据的标准化、图像特征提取）、化学试剂及仪器型号、技术验证等。

（3）数据样本描述。数据期刊投稿指南中规定的数据样本描述是对样本来源、数据结构的总括性描述，以突出实验样本的权威性和可靠性。4 种数据期刊对数据样本描述做出了细节要求。例如，期刊 1 在规格表一栏中要求作者描述：①实验因素，即介绍对样本的预处理过程；②实验特点，简短描述实验过程；③数据收集位置，描述收集样本数据的地理位置（及 GPS 坐标）。期刊 6 在“数据样本描述”模块中，要求具体给出数据集的典型样本，并描述样本的来源、数据结构等信息，以帮助读者理解数据集的内涵，要求作者在选择数据样本时秉持代表性和精简性相结合的原则。与期刊 1 一致，期刊 6 同样提出需对数据样本的地理经纬度、地理区域做出详细描述。

4.2.4　应用场景　数据应用场景指的是数据论文创作者对所描述和介绍的科学数据，在主要用途、技术验证和潜在价值等方面加以说明。通过对所选择的 10 种数据期刊的数据论文投稿指南和写作须知等规范文档的内容分析可知，数据应用场景主要体现在数据可用性、数据价值、数据记录、技术验证 4 个模块中。10 种数据期刊的投稿指南有关数据应用场景的描述要求见表 6。

（1）数据可用性。数据可用性是指在某一特定情景下数据使用者可以重用科学数据以开启或完成某项科研工作的情况说明。包括科学的可获取性、可理解性、使用效率和效果等。经调研，在 10 种数据期刊中，期刊 2、期刊 3、期刊 4 及期刊 7 强制要求在数据论文中提供数据的可用性信息。进一步分析可知，不同数据期刊对数据可用性描述在要求上有所不同。例如，期刊 2 重点提出，要提供数据集生成和处理过程中相关程序的源代码，即“代码可用性”模块，说明用户如何访问和使用这些程序代码，可能受到哪些限制。此外还应提供所应用软件的版本，以及用于生成、测试或处理当前数据集的特定变量或参数信息。

（2）技术验证。技术验证是指为支撑数据集的技术质量所需提供的实验或分析。这可以提高数据（集）的可信度和可靠性，并为科学研究人员重用和复

表 6　10 种数据期刊对数据论文在揭示数据应用场景方面的写作要求

期刊编号 \ 核心模块	数据可用性	技术验证	数据记录	数据价值
1	–	–	–	提供数据对比挖掘、数据共享途径、数据应用方法
2	说明自定义代码的访问途径、访问限制、软件版本、参数信息	验证数据的有关实验、实验误差的统计分析等内容	引用数据应遵循规范化引用格式，标明研究样本、数据来源、实验操作	在背景介绍部分要求简要描述数据潜在重用价值
3	版权限制、专有限制、成本等内容	通过数据质量验证确保数据真实、可靠	–	在摘要部分简要描述数据价值
4	提供数据和材料的可用性声明，数据超链接	–	–	–
5	–	–	–	–
6	数据使用的时间、地点、条件、存储库名称、DOI、许可证等	–	–	覆盖范围、加工方法、潜在应用价值
7	–	提供数据集操作的验证信息，控制数据误差或偏移	–	–
8		–	–	数据集潜在和实际价值
9	–	–	–	–
10	–	–	–	–

用科学数据提供帮助。具体来说，期刊 2 指出，可根据实际需要以图表形式提供技术验证，以证明数据的可靠性，主要内容应包括：验证数据收集程序的实验、有关实验误差的统计分析等内容。期刊 3 和期刊 7 均要求提供用于展示数据集操作的所有验证信息，即如何控制数据误差或偏移，并通过数据质量验证确保数据的真实可靠。

（3）数据记录。数据记录是用于解释和记录与某一科研过程密切相关的全部数据，包括存储信息的公共存储库名称、数据文件及其格式概述。在所调研的 10 种数据期刊中，仅期刊 2 对数据记录模块做出了较为细致的内容规定，要求每条外部数据记录都需遵循特定的引用格式，并完成对数据的引用。同时，采用表格形式对数据记录进行可视化展示，标明研究样本、数据来源、实验操作等信息。若数据记录中包含分析和收集得到的数据，还需对数据输出的过程予以记录。

（4）数据价值。数据价值指的是科学数据的特定属性和功能能够满足数据使用者科研所需的效用和价值。分析发现，有 5 种数据期刊对数据价值模块做出了不同的详细说明（见表 5）。其共性特征包括：提出数据在来源、加工过程、质量控制等方面的创新性，并就数据覆盖范围、加工方法、潜在应用价值等方面说明数据集的价值。

4.2.5　使用方法　数据使用方法的核心内容是为以数据为中心的永久性科学记录提供可靠的用法说明，将数据创建者存储科学数据的位置、数据创建阶段使用的软件包、代码、加工流程和操作步骤、具体规程等一并提交给数据使用者，进而促进科学数据沿着正确的路径被应用、共享或重现。剖析所选择的这 10 种数据期刊的投稿指可知，有关数据使用方法的内容主要体现在数据集存储位置、数据结构描述、数据类型格式与时效性说明以及数据使用说明 4 个模块，见表 7。

（1）数据（集）存储位置。数据论文中描述的数据集存储位置指的是数据的物理存储地点或网络存储空间的位置，如某一地点、公共存储库或第三方存储库等。为方便数据的传播与共享，通常需要在数据论文中提供数据存储位置的超链接，并与所描述的数据集相关联。所调研的 10 种数据期刊中有一半的数据期刊强制性要求数据论文提供数据存储位置的描述模块（见表 7）。5 种数据期刊要求在数据论文中提供数据集存储的永久性标识符（如 DOI 或 URL），以促进数据集的快速获取和利用。

（2）数据结构描述。数据结构描述是数据论文对科学数据的变量、数值等细节性问题的揭示。除期刊 9 外，所选取的 9 种数据期刊均在其投稿指南中提及数据结构的描述信息，明确提出应当提供产生、测试和处理数据集的所有变量和参数。其中，期刊 3 对数据结构描述更为详尽，要求涉及的内容包括：①变量的特征、定义、计量单位、存储类型、列表、变量范围、缺失值、精度、格式等；②缺失数据、异常数据、标准误差等。

（3）数据格式、类型与时效性说明。数据格式、类型与时效性说明是数据论文对科学数据在引用、重用

表7 10种数据期刊对数据论文在呈现数据使用方法方面的写作要求

期刊编号＼核心模块	数据(集)存储位置	数据结构描述	数据格式、类型与时效性说明	数据使用说明
1	-	数据集中提供变量和样本信息	要求提供数据类型和数据格式	-
2	-	提供产生、测试或处理数据集的特定变量和参数	提供数据文件及格式概述	数据分析软件包、数据处理建议、代码
3	数据存储位置和介质	变量的特征、计量单位、列表、取值范围、精度、格式等、缺失数据、异常数据、标准误差等	要求对数据状态予以说明:数据更新时间、存档时间、元数据状态等	说明引用数据方法
4	数据存储位置(公共存储库、标识符)	说明缺失值、小样本数据等情况	将相同类型数据归并为数据集	-
5	存储位置(如URL),将数据集与DOI编号相关联	说明每一项数据所包含的内容	数据类型、结构、数据集许可	提供数据集使用说明的详细介绍
6	数据服务系统网址	举例说明多种数据结构的数据集内容	要求提供数据量、数据格式的具体信息	推荐软件、数据加工方法、数据集成和对比说明
7	-	提供缺失值、变量、参数	使用SAV、SAS或POR等数据格式	-
8	-	数据参数、变量信息	数据类型、版本、格式	-
9	数据存储位置、DOI	-	要求提供数据的格式(如ASCLL、CSV、SPSS、SAS)、版本、出版时间、类型、数据集许可	数据聚合、分析、引用、验证等
10	-	数据的物理结构、异常值、缺失值、附加信息	使用通用型数据格式、提供数据集许可	-

时需要考虑的关键要素进行系统描述的模块。这一模块的内容涉及科学数据的创建时间、科学数据的首次使用时间、语言、版本、数据规模、数据类型、数据格式和数据集许可等。所调研的10种数据期刊的投稿指南均对这些内容提出了明确的要求。

(4)数据使用说明。数据期刊投稿指南中规定的数据使用说明是指数据期刊要求数据论文创作者介绍科学数据的使用方法,以更好地促进科学数据的引用、分析、验证或重用等。总结来看,有5种数据期刊要求数据论文提供数据使用说明。这些数据期刊对数据使用说明的内容规定有一些共性,如要求包含:数据的使用方法和建议、生成数据的软件或代码、数据处理流程和具体步骤以及其他补充材料等。不过,期刊9还从格式方面对数据使用说明进行了描述,要求提交50-200个单词的数据使用和重用方法。

4.2.6 补充说明 除前述的五个内容模块,所调研的10种数据期刊均在投稿指南设置有补充说明。整体上看,补充说明主要包括补充信息、作者贡献、致谢和参考文献4项内容,见表8。

(1)补充信息。补充信息是数据论文中用来声明与科学数据相关的专利、利益冲突、注释、缩略词表、相关研究文章等内容的模块。其中,①在利益冲突声明方面,有5种数据期刊要求公开发表与科学数据相关的利益声明(见表8);②在专利方面,期刊6提出,若在科学数据创建过程中获得了相关的专利成果,可以添加专利信息予以说明,但此部分并非强制性要求;③在附录方面,期刊5认为,附录同样是非强制性的内容模块,可用于补充数据论文涉及的其他必要信息,对理解和再现科学实验过程至关重要;④在缩略语方面,期刊1和期刊4要求在数据论文首页采用脚注形式定义所有专业词汇的缩写形式,确保数据论文中出现的缩略语表达一致;⑤在脚注方面,期刊1和期刊4要求,数据论文中出现的脚注应连续编号,并在结尾处单独列出,但是应减少脚注的使用;⑥关于图、表等的说明,这部分内容主要包括图表格、数学公式、视频等文件、单位和符号等的说明,用于概述生成和分析数据的输出过程。

(2)作者贡献。作者贡献模块用于详细列出每位作者对创建和描述数据所做出的贡献,以便为解决科学数据的归属、质量和引用等问题提供依据。经调研,有6种数据期刊明确规定需要在数据论文的结尾处具体列出每位作者承担的任务。其中,期刊7对作者贡献部分的要求最为细致,列举了包括模型构建、数据管理、形式化分析、调查、方案设计、项目管理等在内的14个项目作为可分配给各位作者的任务。

(3)致谢。致谢是数据论文创作者用来向相关个人表示感谢的内容模块。在技术支持、数据整理、写作指导、翻译校对等方面为数据论文创作者提供支持的人员,都可以在这个模块中列出。总体来看,所有数据期刊都重视在数据论文中提供致谢信息,通常会要求

表 8 10 种数据期刊对数据论文中补充说明的写作要求

期刊编号 \ 核心模块	补充信息	作者贡献	致谢	参考文献
1	利益声明，披露与数据相关的多方利益关系；采用脚注形式标明缩略语；脚注连续编号，单独列出	–	对数据论文创作提供帮助相关主体表示感谢	列出相关参考资料
2	若有利益冲突则提供利益声明	结尾处详细描述	简短介绍对研究做出贡献的人员	采用标准格式引用
3	–	–	在致谢中说明每位作者的具体贡献	–
4	清楚解释与数据相关的财务利益冲突；提供缩略词表	标明每位作者贡献	对为数据创作提供援助的作者致谢	根据不同文献类型格式完成列出参考文献
5	识别和声明利益冲突；提供附录，用于补充和解释实验过程	分段具体列出每位作者承担的任务	感谢其他人员提供的技术、管理支持	依次列出参考文献
6	说明数据创建中生成的专利	简要描述每位作者贡献	感谢为数据论文创作做出贡献的其他作者	使用统一编号和格式依次列出
7	附加财务、个人等利益冲突	从 14 个项目详细描述作者贡献的具体内容	感谢对数据论文研究和写作过程做出贡献的作者	无强制性引用格式要求，与引用顺序一致
8	–	–	对数据论文创作做出贡献的作者致谢	根据不同出版物类型采用不同格式
9	数据收集方法描述（100 – 200 个单词），影响因素、项目数量、操作过程	列出每位作者具体贡献	向未直接参与数据处理的人员致谢	以 APA 格式输入参考文献
10	试验设计、测量步骤、化学试剂和仪器	提交作者名单，说明作者贡献	列出不符合作者要求但提供支持的个人	–

数据论文创作者简洁地列出参与数据采集或提供其他技术帮助的科研人员。

（4）参考文献。与传统研究性论文类似，数据论文也需要对已有的相关研究成果进行借鉴和参考，并采用统一的格式和编号加以引用。除期刊 3 和期刊 10 未提出强制性要求之外，其他 8 种数据期刊都要求在数据论文的结尾处按照出现在数据论文正文中的顺序依次列出参考文献，并对含有 DOI 或 URL 的内容加以标注。

5 讨论与结论

通过上述分析可以发现，当前在科学数据出版界较为活跃的 10 种数据期刊所规定的数据论文的内容框架，具有以下特点：

（1）并没有一种数据期刊的数据论文投稿指南或写作须知能够全部涵盖数据论文的所有模块。从表 1 可以看出，即便像《数据简述》（*Data in Brief*）这种已出版数据论文达到 3 000 多篇的数据期刊，也只是相对较多地纳入数据论文的通用模块，对于用来描述科学数据完整性、数据（集）优劣程度和应用场景等方面的内容，目前并没有明确涉及。而且多数数据期刊对数据论文涉及的学科范围也没有进行规定，事实上，科研人员在选择科学数据，或者是借助数据论文来选择科学数据时，更会优先考虑科学数据或数据论文的主题相关性。当然，由于不同数据期刊的侧重存在一定的差异性，事实上也难以出现一个全面的模块框架。

（2）数据论文内容构成与传统学术论文的内容构成既有联系又有区别。目前，数据论文的内容构成主要可以分为两个部分：一部分是与数据集相关的描述款目，另一部分是与传统科技论文相类似的描述款目。其中，数据集相关的描述款目是指数据论文独有的、有别于传统科技论文的款目，是数据论文赋予科学数据被正确描述、引用和重用的重要属性，这类款目涵盖数据质量评估、数据产生背景、获得方法和实验方法、使用方法和应用场景等。而与传统科技论文相类似的描述款目，主要有标题、作者、摘要、关键词、参考文献等。需要指出的是，尽管多数数据期刊投稿指南中明确要求数据论文应当包含上述传统科技论文相类似的描述款目，但进一步分析可知，这些描述款目虽然使用的名称一样，但通常更要在内容上符合科学数据的特性、符合数据论文的功能定位。比如，部分数据期刊要求数据论文的标题中含有“数据”“数据集”等字样，而摘要只需概述数据来源、实验方法和研究设计等，不需要像传统科技论文的一样，摘要还需要给出主要的研究发

现。

（3）数据论文内容框架的必备模块侧重于描述科学数据的突出特性。总体来看，当前刊载数据论文数量排名前10位的数据期刊，在其投稿指南或写作模板中，要求作者全面且详尽描述的内容模块主要涉及标题、摘要、作者、实验设计，获得方法和实验方法，以及使用方法中的数据结构描述、数据格式、类型与时效性说明等。同时，这些核心模块也要求重在描述科学数据的特点。例如，多数数据期刊要求数据论文的标题应揭示体现科学数据的来源及其使用价值；摘要需侧重对研究目的、数据产生背景、实验条件及方法、数据价值等内容进行概括性描述，而详细内容介绍应放在数据论文主体部分－－数据产生背景、获得方法和实验方法，以及使用方法中。同时，要求明确创建数据论文的科研人员即数据论文的作者，若不具备作者资格、但对数据论文的创作做出贡献的相关人员则应该在致谢中加以说明。

参考文献：

［1］黄国彬，屈亚杰，王舒．UKDA和ICPSR社科数据发布平台数据管理功能剖析［J］．图书情报工作，2017，61(21)：40－48.
［2］黄国彬，刘馨然，姜颖．影响科学数据引用的外部因素分析［J］．数字图书馆论坛，2017(6)：2－8.
［3］刘凤红，张恬．开放科学背景下新兴学术论文出版类型——研究要素出版［J］．中国科技期刊研究，2017，28(2)：138－144.
［4］刘凤红，崔金钟，韩芳桥，等．数据论文：大数据时代新兴学术论文出版类型探讨［J］．中国科技期刊研究，2014，25(12)：1451－1456.
［5］田稷，陈珲夏．数据期刊及数据论文定量分析与思考［J］．图书馆论坛，2016，36(3)：42－48.
［6］黄如花，李楠．基于数据生命周期模型的国外数据期刊政策研究［J］．图书与情报，2017(3)：36－42，108.
［7］李修．数据期刊出版质量控制研究［J］．编辑之友，2017(4)：33－38.
［8］CANDELA L，CASTELLI D，MANGHI P，et al. Data journals：a survey［J］. Journal of the Association for Information Science and Technology，2015，66(9)：1747－1762.
［9］ROA－MARTINEZ S M，VIDOTTI S A B，SANTANA R C. Proposed structure of a data paper structure as scientific publication［J/OL］. Revista espanoõla de documentacioón cientifica，2017，40(1)：e167.［2019－08－09］. http://dx.doi.org/10.3989/redc.2017.1.1375.
［10］屈宝强，王凯．数据论文的出现与发展［J］．图书与情报，2015(5)：1－8.
［11］CHAVAN V，PENEV L. The data paper：a mechanism to incentivize data publishing in biodiversity science［J］. BMC bioinformatics，2011，12(6)：2399－2405.
［12］Dryad. Look up your journal［EB/OL］.［2019－01－07］. https://datadryad.org/pages/journalLookup.
［13］Data in Brief. Guide for authors［EB/OL］.［2019－01－17］. https://www.elsevier.com/journals/data－in－brief/2352－3409/guide－for－authors.
［14］Scientific Data. Submission guidelines［EB/OL］.［2019－01－17］. https://www.nature.com/sdata/publish/submission－guidelines.
［15］Ecology. Data papers［EB/OL］.［2019－01－17］. https://esajournals.onlinelibrary.wiley.com/journal/19399170.
［16］BMC Research notes. Submission guidelines［EB/OL］.［2019－01－17］. https://bmcresnotes.biomedcentral.com/submission－guidelines/preparing－your－manuscript/data－note.
［17］Data. Instructions for authors［EB/OL］.［2019－01－17］. https://www.mdpi.com/journal/data/instructions.
［18］中国科学数据．投稿须知［EB/OL］．［2019－01－17］．http://www.csdata.org/p/static/143/.
［19］F1000research. Article guidelines［EB/OL］.［2019－01－17］. https://f1000research.com/for－authors/article－guidelines/data－notes.
［20］Geoscience data journal. Author guidelines［EB/OL］.［2019－01－17］. https://rmets.onlinelibrary.wiley.com/hub/journal/20496060/about/author－guidelines.
［21］Journal of open psychology data author guidelines［EB/OL］.［2019－01－17］. https://openpsychologydata.metajnl.com/about/submissions/.
［22］Ecological research. Data papers template.［EB/OL］.［2019－01－17］. http://www.greynet.org/thegreyjournal/datapapertemplate.html.

作者贡献说明：

黄国彬：负责论文的选题确定，论文的撰写、修改与指导；
郑霞：负责论文的素材收集与整理，论文的撰写与投稿。

Research on Content Normalization of Data Paper

Huang Guobin　Zheng Xia
School of Government，Beijing Normal University，Beijing 100875

Abstract：［**Purpose/significance**］The standardized management，citation and reuse of scientific data has attracted

wide attention in academic circles. In this context, a large number of data papers and data journals have emerged to promote the rational use of scientific data. However, data paper has not yet formed a unified and standardized system. The study will summarize and refine the content framework and core modules of data paper published by data journals which are active in the research data publishing field, in order to provide reference for relevant personnel or organizations in writing, use and management of data papers. [**Method/process**] Comprehensive application of content analysis and comparative research methods, based on the thinking mode of scientific data users in searching, acquiring and reusing research data, and based on the submission guides or writing instructions issued by different data journals, the content framework and core modules of data paper were analyzed from six perspectives, including subject relevance, data quality, data generation and acquisition methods, application scenarios, usage methods and supplementary instructions. [**Result/conclusion**] At present, there is no submission guide or writing instructions of data journals that can fully cover all modules of data paper; the content composition of data paper is related to and different from that of traditional academic paper; the essential module of content framework of data paper focuses on describing the prominent characteristics of scientific data; a normalized description framework for features of scientific data is still on the way.

Keywords: data paper scientific data content framework content normalization